JOURNAL OF VIROLOGY, Sept. 1996, p. 6083–6096 Vol. 70, No. 9
0022-538X/96/$04.00+0

A Reevaluation of the Higher Taxonomy of Viruses Based on RNA Polymerases

PAOLO M. DE A. ZANOTTO,[1] MARK J. GIBBS,[1]† ERNEST A. GOULD,[1] AND EDWARD C. HOLMES[2]*

Natural Environment Research Council, Institute of Virology and Environmental Microbiology, Oxford OX1 3SR,[1] and Wellcome Center for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford, Oxford OX1 3PS,[2] United Kingdom

Received 22 November 1995/Accepted 4 June 1996

In order to assess the validity of classifications of RNA viruses, published alignments and phylogenies of RNA-dependent RNA and DNA polymerase sequences were reevaluated by a Monte Carlo randomization procedure, bootstrap resampling, and phylogenetic signal analysis. Although clear relationships between some viral taxa were identified, overall the sequence similarities and phylogenetic signals were insufficient to support many of the proposed evolutionary groupings of RNA viruses. Likewise, no support for the common ancestry of RNA-dependent RNA polymerases and reverse transcriptases was found.

RNA viruses can be placed into four main categories based on their replication and coding strategies: positive- and negative-strand RNA (+RNA and −RNA), double-stranded RNA (dsRNA) viruses, and retroviruses. The genomes of the largest RNA viruses are about 3×10^4 bases long, and many have had their complete nucleotide sequences determined (4, 10). Analysis of these sequence data has revealed extensive diversity among the viruses in each of the four types: viruses from different viral families usually have different genome organizations, often with nonhomologous genes, and employ distinct strategies of gene expression.

Using sequence data to group RNA virus isolates into species, genera, and families is usually not controversial. However, support for groupings higher than the family level is more circumspect because of a lack of shared characteristics. One of the few features all RNA viruses, with the exception of retroviruses, have in common is that they encode RNA-dependent RNA polymerases (RdRp) which are employed in replication. The sequences of the RdRp genes are among the most conserved from these viruses (14). Recently the RdRp sequences of viruses from a wide range of genera and families have been used to infer phylogenies that cluster together most or all +RNA, −RNA, and dsRNA viruses. These phylogenies have formed the basis for various classification schemes for these viruses above the family level (6, 7, 10, 14, 23). Koonin (13) and Koonin and Dolja (14) propose three supergroups of +RNA viruses from which dsRNA viruses originated on separate occasions. Each of these three supergroups (*Picornavirata*, *Flavivirata*, and *Rubivirata*) is considered a class, with subsequent divisions into orders, families, and genera or groups (14). This scheme, supported by the observation of the same three supergroups in helicase sequences (14), was also postulated by Dolja and Carrington (6) and Ward (22). However, different RdRp phylogenies have been reported. Goldbach and de Haan (10) produced a phylogenetic tree which implies that the first split was between dsRNA viruses and +RNA viruses and that both segmented and nonsegmented −RNA viruses then evolved from a +RNA virus lineage. Bruenn (5) presents a partially resolved tree in which the leviviruses (a phage lineage) do not belong in the supergroup II of Koonin and Dolja (14) but were classed as an outgroup of the +RNA viruses. Bruenn's tree also implies that a +RNA lineage split into picornaviruses and dsRNA viruses and that all insect-associated +RNA viruses originated from dsRNA viruses. Comparisons of RdRp sequences with those from RNA-dependent DNA polymerases (the reverse transcriptases [RT]) have also suggested that RdRp and RT are related proteins, although this is only based on the colinearity and conservation of four sequence motifs shared between them (18).

Given the striking disagreements among the evolutionary hypotheses reported in the literature, it is surprising that the suitability of RdRp as a keystone phylogenetic marker has not been rigorously questioned. Here we undertake such an analysis, which should precede viral classifications and any subsequent nomenclature. Although we concentrate on phylogenies inferred from RdRp sequences, we will also address the claim that the RdRp shares a common ancestor with the RT of retroviruses.

MATERIALS AND METHODS

Sequence data. Four aligned amino acid data sets, all of which were taken from the literature, were used in this analysis: (i) an 80-residue-long data set incorporating five highly conserved sequence motifs from 40 RdRp and 40 RT by Poch et al. (18); (ii) a 520-residue-long data set representing 50 RdRp by Bruenn (5); (iii) a 380-residue-long data set of 46 RdRps by Koonin (13); and (iv) a 120-residue-long data set, including the 50 RdRp presented by Koonin and Dolja (14), with the addition of 9 other RdRp from Koonin (13). This new data set adhered strictly to the alignment scheme proposed by Koonin and Dolja (14). The number of RdRp sequences currently available is larger than those analyzed here, but we have chosen to examine data sets and alignments which have been previously used in viral classification. A full list of the viruses (and other agents) from each of the four data sets is given in Table 1. In all cases we have kept the same virus abbreviations as those employed in the original publications so that our results could be compared directly.

Monte Carlo analysis. Sequence similarity was checked against a null hypothesis of randomness by using a Monte Carlo randomization test (16) incorporated in the MULTIALIGN program (3). The program was used to generate 100 random sequences of the same length and amino acid composition as each of the original sequences. The original sequences were then aligned with each randomized sequence to generate a distribution of similarity scores. Sequence similarity was then measured as the number of standard deviations (SD) above the mean value observed for the random-sequence data set. SD values (i.e., the numbers of

* Corresponding author. Mailing address: Wellcome Center for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford, South Parks Rd., Oxford, OX1 3PS, United Kingdom. Phone: 44 1865 271282. Fax: 44 1865 310447. Electronic mail address: Edward.Holmes@zoo.ox.ac.uk.

† Present address: Co-Operative Research Center for Plant Science, c/o CSIRO Division of Plant Industry, Canberra 2601, Australia.

TABLE 1. List of agents used in this study[a]

Agent(s)	Abbreviation(s) used by:			Family or group
	Poch et al. (18)	Bruenn (5)	Koonin and Dolja (14)	
RT				
Human hepatitis B virus	HepB			*Hepdnaviridae*
Woodchuck hepatitis B virus	HepWo			*Hepdnaviridae*
Duck hepatitis B virus	HepBDu			*Hepdnaviridae*
Human endogenous retrovirus C	HERVC			Retrovirus
AKV murine leukemia virus	AKVLMV			Retrovirus
Murine Moloney leukemia virus	MoMLV			Retrovirus
Hamster intracisternal A particle	IAPH18			Retrovirus
Rous sarcoma virus	RSV			Retrovirus
Simian Mason-Pfizer monkey virus	SMPV			Retrovirus
Murine mammary tumor virus	MMTV			Retrovirus
Human endogenous retrovirus K	HERVK			Retrovirus
Human adult T-cell leukemia virus	ATLV			Retrovirus
Human T-cell leukemia virus type 2	HTLVII			Retrovirus
Bovine leukemia virus	BLV			Retrovirus
Human immunodeficiency virus type 1	HIV 1			Retrovirus
Human immunodeficiency virus type 2	HIV 2			Retrovirus
Caprine arthritis encephalitis virus	CAEV			Retrovirus
Equine infectious anemia virus	EIAV			Retrovirus
Visna virus	Visna			Retrovirus
Drosophila 17.6 element	17.6			Gypsy-like
Drosophila 297 element	297			Gypsy-like
Drosophila gypsy element	Gypsy			Gypsy-like
Drosophila 412 element	412			Gypsy-like
Cauliflower mosaic virus	CaMV			Gypsy-like
Dictyostelium DIRS-1 element	Dirs			Gypsy-like
Ty912 element	TY912			Ty-like
Drosophila 1731 element	1731			Ty-like
Drosophila copia element	Copia			Ty-like
Mauriceville plasmid (mtDNA)[b]	MauP			Line-like
Chlamydomonas intron (mtDNA)	RTChla			Line-like
Trypanosoma ingi element	Ingi			Line-like
Drosophila F factor	Ffac			Line-like
Maize Cin4 element	CIN4			Line-like
Drosophila 1 factor	Ifac			Line-like
Yeast class I intron (mtDNA)	IntSp			Line-like
Yeast class I introns (mtDNA)	Int31, Int32			Line-like
Mouse line-1 element	LiMd			Line-like
Prosimian and human line-1 elements	LIS1, L1Hu			Line-like
RdRp				
Bacteriophage MS2	MS2V	Ms2	MS2	*Leviviridae*
Bacteriophage Ga	GaV	Ga	GA	*Leviviridae*
Bacteriophage Qβ	QBeta V	Qbeta	QBETA	*Leviviridae*
Bacteriophage SP		Sp	SP	*Leviviridae*
Bacteriophage ϕ6		Phi6		*Leviviridae*
Poliovirus	PolV	Polio	PV	*Picornaviridae*
Coxsackievirus	CoxV	Coxv		*Picornaviridae*
Human rhinovirus type 14	HRV14	Hrv14		*Picornaviridae*
Human rhinovirus type 2	HRV2			*Picornaviridae*
Encephalomyocarditis virus	EMCV	Emc	EMCV	*Picornaviridae*
Foot-and-mouth disease virus	FMDV	Fmdv	FMDV	*Picornaviridae*
Echovirus 22			ECHO 22	*Picornaviridae*
Hepatitis A virus	HAV	Hav	HAV	*Picornaviridae*
Feline calicivirus			FCV	Calicivirus
Rice tungro spherical virus			RTSV	Waika
Hungarian grapevine chrome mosaic virus			GCMV	Nepovirus
Tomato black ring virus	Tbrv			Nepovirus
Cowpea mosaic virus	CPMV	Cpmv	CPMV	*Comoviridae*
Southern bean mosaic virus			SBMV	Sobemovirus
Black beetle virus	BBV	Bbv	BBV	*Nodaviridae*
Tobacco etch virus	TEV	Tev	TEV	Potyvirus
Tobacco vein mottle virus	TVMV	Tvmv		Potyvirus
Plumpox virus		Ppv		Potyvirus
Pepper mottle virus			PEMV2	Potyvirus
Theiler's murine encephalomyelitis virus	TMEV	Tmev		*Picornaviridae*
Sindbis and Middleburg viruses	SinV, MidV	Sinv	SNBV	*Togaviridae*

Continued on following page

TABLE 1—*Continued*

Agent(s)	Abbreviation(s) used by:			Family or group
	Poch et al. (18)	Bruenn (5)	Koonin and Dolja (14)	
Semliki Forest virus	SFV	Sfv		*Togaviridae*
O'nyong-nyong virus		Onv		*Togaviridae*
Ross River virus		Rrv		*Togaviridae*
Pea enation mosaic virus			PEMV2	PEMV
Tobacco mosaic virus	TMV	Tmv	TMV	Tobamovirus
Beet necrotic yellow vein virus	BNYVV	Bnyvv	BNYVV	Furovirus
Mouse hepatitis virus		Mhv		*Coronaviridae*
Infectious bronchitis virus			IBV	*Coronaviridae*
Berne virus			BEV	Torovirus
Barley stripe mosaic virus		Bsmv	BSMV	*Hordeiviridae*
Equine arteritis virus			EAV	Arterivirus
Red clover necrotic mosaic virus			RCNMV	Dianthovirus
Potato virus X		Pvx	PVX	*Potexviridae*
Brome mosaic virus	BMV	Bmv	BMV	*Tricornaviridae*
Tobacco rattle virus	TRV	Trv	TRV	Tobravirus
Alfalfa mosaic virus	AaMV	Almv	ALMV	*Tricornaviridae*
Cucumber mosaic virus	CucMV	Cucmv	CMV	*Tricornaviridae*
Carnation mottle virus	CarMV	Carmv	CarMV	Carmovirus
Maize chlorotic mottle virus			MCMV	Carmovirus
Turnip crinkle virus			TCV	Carmovirus
Turnip yellow mosaic virus	TYMV	Tymv	TYMV	*Tymoviridae*
Barley yellow mosaic virus			BaYMV	Bymovirus
Barley yellow dwarf virus	BYDV	Bydv	BYDV	Luteovirus
Potato leaf roll virus		Plrv	PLRV	Luteovirus
Beet western yellow virus		Bwyv	BWYV	Luteovirus
Tomato bushy stunt virus			TBSV	Tombusvirus
Cucumber necrosis virus		Cnv	CNV	—[c]
Cymbidium ringspot virus		Cyrv		Tombusvirus
Maize chlorotic mottle virus		Mcmv		*Bromoviridae*
Yellow fever virus	YFV	Yfv	YFV	Flavivirus
Dengue virus serotype 4		Dengue	DEN4	Flavivirus
West Nile virus	WNV	Wnv	WNV	Flavivirus
Japanese encephalitis virus		Jev		Flavivirus
Tick-borne encephalitis virus			TBEV	Flavivirus
Cell fusion agent			CFAV	Flavivirus
Hepatitis C virus			HCV	Flavivirus
Hepatitis E virus			HEV	—
Rubella virus			RubV	Rubivirus
Apple chlorotic leafspot virus (ACSLV)			ACLV	—
Apple stem grooving virus			ASGV	Capillovirus
Infectious bursal disease virus	IBDV	Ibdv		*Birnaviridae*
Bluetongue virus	BTV	Btv		*Reoviridae*
Influenza A and B viruses	InfA, InfB			*Paramyxoviridae*
Tacaribe virus	TacaV			*Arenaviridae*
Lymphocytic choriomeningitis virus	LCMV			*Arenaviridae*
Newcastle disease virus	NDV			*Paramyxoviridae*
Sendai virus	SendV			*Paramyxoviridae*
Measles virus	MeasV			*Paramyxoviridae*
Rabies virus	RabV			*Rhabdoviridae*
Vesicular stomatitis virus	VSV			*Rhabdoviridae*
Saccharomyces cerevisiae virus L1		Scv11		Totivirus
Saccharomyces cerevisiae virus La		Scvla		Totivirus
Bovine rotavirus		Rot		*Reoviridae*
Reovirus		Reo		*Reoviridae*
Cryphonectria parasitica hypovirulence virus			dsHyAV	dsRNA virus
Saccharomyces cerevisiae virus L-A			dsScV	dsRNA virus
Leishmania RNA virus 1			dsLRV1	dsRNA virus
Saccharomyces cerevisiae W RNA			WRNA	dsRNA virus
Saccharomyces cerevisiae T RNA			TRNA	dsRNA virus

[a] The abbreviations used in the original publications were kept so that our results could be compared directly with them. The data set of Koonin (13) was excluded because the abbreviations used here are also given in the work of Koonin and Dolja (14).
[b] mtDNA, mitochondrial DNA.
[c] —, unassigned.

SD above the mean values) for each pairwise comparison are presented as density plots on the grey scale from white (maximum similarity, high SD value) to black (minimum similarity, low SD value). The density plots were constructed by using Mathematica version 2.2 (Wolfran Research, Inc.).

Bootstrap analysis. It is also informative to determine whether phylogenetic trees reconstructed from the four data sets provide a consistent picture of polymerase relationships. The robustness of a series of phylogenetic trees reconstructed by a variety for methods was therefore assessed by the bootstrap re-

sampling method. For each data set, 100 resampled data sets were generated with the SEQBOOT program in PHYLIP 3.5 (8). Distance matrices were generated with the PHYLIP PROTDIST program with the "categories" distance model, which accounts for the chemical similarity of the protein sequences and is appropriate for sequences as divergent as the RdRp. Phylogenetic trees were constructed from the distance matrices by the Fitch-Margoliash, neighbor-joining, and UPGMA clustering methods (PHYLIP programs FITCH and NEIGHBOR). Additionally, bootstrap analyses were conducted by the parsimony methods in PAUP 3.1 (20) and with the PROTPARS program in PHYLIP 3.5. Strict, semistrict, and 50% majority rule consensus trees were calculated by using PAUP 3.1. Two amino acid-weighting schemes were used in the parsimony analysis, (i) one in which the number of amino acid changes is set by the minimum number of nucleotide substitutions determined by the genetic code and (ii) one in which amino acids are grouped according to physicochemical similarity on the basis of the schemes proposed by Koonin (Ala and Gly; Leu, Ile, Val, and Met; Phe, Tyr, and Trp; Ser and Thr; Asn, Gln, Asp, and Glu; Arg and Lys; Cys; Pro; and His) (13) and Bruenn (Gly and Pro; Ile, Met, Cys, Leu, Ala, and Val; Tyr, Trp, Phe, and His; Thr and Ser; Asn, Gln, Asp, and Glu; and Arg and Lys) (5).

Phylogenetic signal analysis using g1 statistics. Additional evidence for a clear phylogenetic signal was sought by calculating the distribution of lengths of all possible bifurcating parsimony trees. Data sets with phylogenetic signal produce skewed tree-length distributions (12, 20). Skewness can be measured by the g1 statistic, and data sets with a phylogenetic signal have g1 values significantly less than 0 (left skew) (8). Because each of the four data sets includes a large number of sequences, we chose, for computational reasons, to infer trees using only a subset of sequences representing the higher-order groupings and deeper nodes from the complete phylogenetic tree (the "tree backbone"). Consequently, we chose a single representative of each "subtree" found in the bootstrap analysis (i.e., clusters of taxa supported more than 50% of the time), to give a maximum of nine taxa. The distribution of parsimony tree lengths was calculated by using PAUP 3.1. This rationale is justified, because phylogenetic signal should be obtained from any subset of taxa from a complete tree. However, random data sets can sometimes yield tree-length distributions with a left skew and the value of g1 has been shown to depend on the number of taxa and the lengths of the sequences used in the analysis (12). Therefore, to demonstrate that the g1 values from the four RdRp data sets could not be obtained by chance, g1 values were also determined from randomized data sets. This was done by generating 30 random data sets, each with the same length and amino acid composition as each original data set, but with no phylogenetic structure, by using the SEQBOOT program. Since the g1 statistic is normally distributed (19), a sample size of 30 constitutes a valid compromise between computational effort and proper estimation of associated error.

Phylogenetic signal analysis using random tree-length distributions. Because exact solutions for measuring the skewness of tree-length distributions for large numbers of taxa are impractical, the minimal (i.e., most parsimonious) tree lengths for larger data sets were calculated and compared with the minimal tree lengths obtained from random data sets. If phylogenetic signal is present in the data, then the RdRp should produce significantly shorter trees than those obtained from random data sets. For a subset of taxa from each of the four data sets, 100 random data sets were generated with the SEQBOOT program. Parsimony trees for each random data set were then reconstructed with the PROTPARS program, and their lengths were used to construct confidence intervals for the length of the minimal random tree by using the SPSS 4.0 statistical package (SPSS, Inc.).

RESULTS

Monte Carlo simulations. To visualize the results of the Monte Carlo simulations for each of the four data sets, levels of sequence similarity, quantified as the number of SD above the random expectation, are presented as density plots (see Fig. 1 to 4). Here SD values are grouped in different shades of the grey scale representing decreasing numbers of SD from white to black. SD values of around 3.0 have been associated with a 1 in 1,000 probability of a match being spurious and were thus considered to signify related sequences (9). However, an analysis of alignments from structurally unrelated proteins resulted in mean values of 3.2 (3), suggesting that this value is too low, especially as values of around 7.55 have been obtained for proteins with completely different secondary structures (2). Furthermore, the SD values presented refer to scores from gap-optimized alignments. This causes the distribution of similarity scores to depart from normality and inflate the levels of sequence similarity (2). Nevertheless, SD values approaching zero are a clear indication of lack of support for sequence relatedness.

The concentration of white and light grey, mostly along the diagonal in the density plots, indicates that for all four data sets, high SD values were mainly obtained among viruses already known to be closely related (within virus families) and not across the whole data set (between virus families). For example, although SD values of greater than 7.5 were obtained for comparisons of the retroviral RT to the transposon sequences from the Poch et al. data set (Fig. 1), suggesting that they are related, there was no similarity between the RT and RdRp sequences (SD value < 2.5) nor between most of the RdRp from different virus families, as indicated by the darkest quadrants.

The results for the Bruenn alignment are presented in Fig. 2. Again, there is little sequence similarity between different virus families. Although the SD values suggest that the RdRp of nepovirus, comovirus, and picornavirus are related, as are the RdRp of tobacco mosaic virus and alphaviruses, it is not possible to infer relatedness between all RdRp of a luteovirus-like group nor within a tobamovirus-like group, where the mouse hepatitis virus showed little similarity to any other group member.

Figure 3 shows the SD values for the Koonin data set. Here SD values for comparisons within representatives of the 11 groups of putative supergroup I (13), (the picorna-, como-, nepo-, poty-, bymo-, noda-, sobemo-, luteo-, corona-, toro-, and nodavirus groups) reached more than 8.34, as indicated by the white regions. However, values for comparisons between members of these groups were low, as indicated by the darker off-diagonal quadrants. Furthermore, nodavirus showed low SD values in comparison with all other members of this supergroup, with the highest value (SD value = 3.51) in comparison with comoviruses. In Koonin's putative supergroup II (the carmo-, tombus-, maize chlorotic mottle, diantho-, luteo-, pesti-, hepatitis C, flavi-, and phage virus groups), low SD values characterized comparisons between phage virus RdRp and the other RdRp as well as between the flavivirus RdRp and the RdRp from the group comprising the carmo-, tombus-, luteo-, and dianthoviruses. The association of pesti- and hepatitis C viruses with the other members of Koonin's supergroup II is also marked by relatively low SD values (mostly below 4.89).

The density plot in Fig. 4 shows the scores obtained from the modified Koonin and Dolja data set. Although some higher SD values can be observed when comparing the putative supergroup II viruses, the off-diagonal values did not increase considerably compared with those in other data sets. Overall, the pattern was similar to that observed for the Koonin data set, indicating a lack of extensive sequence similarity between the RdRp of RNA viruses.

Bootstrap analysis. The results of the bootstrap resampling analysis for the four RdRp data sets are also given (see Fig. 5 to 8). The strict and semistrict consensus methods resulted in completely unresolved trees for the six phylogenetic methods used, implying a lack of congruent phylogenetic signal. The less stringent majority rule consensus method indicated some regions of agreement, and these results are presented here.

There is a striking lack of support for interfamily associations in any of the data sets. In Fig. 5 the bootstrap trees for the Poch et al. data set indicated that only a few subtrees, including usually no more than five taxa, could be resolved with more than 50% bootstrap support. No clades which included both RT and RdRps were obtained. The bootstrap values for the Bruenn data set are shown in Fig. 6. As with the Poch et al. data, parsimony trees showed some highly supported subtrees, but these were not obtained consistently and were also dependent on the model of amino acid replacement used. The picornaviruses (around 70% bootstrap support), the togaviruses (alphaviruses) (100%), the phages (around 90%), and the po-

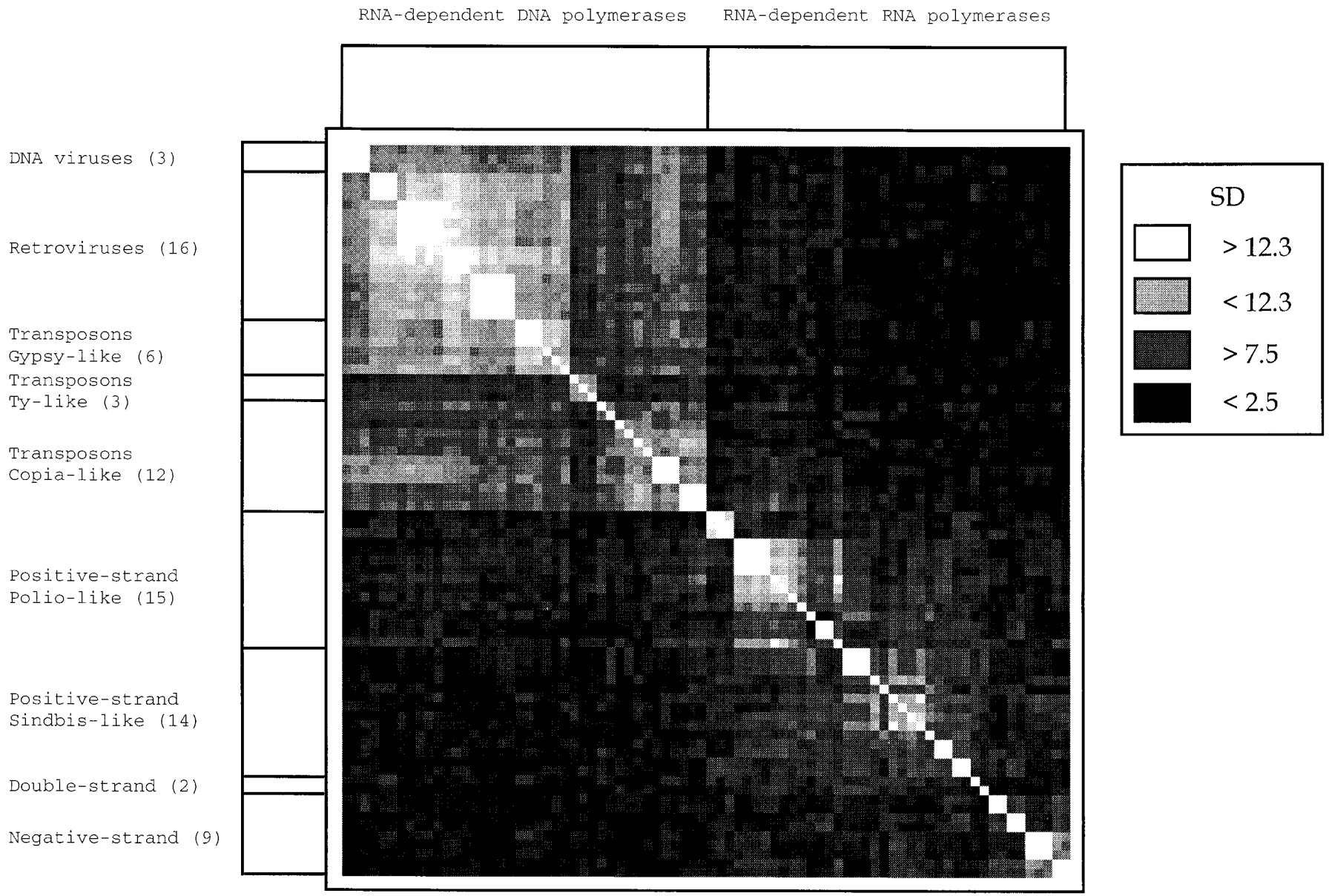

FIG. 1. Monte Carlo simulation analysis of sequence similarity among 40 RdRp and 40 RT from Poch et al. (18). For each pairwise comparison, the number of SD above the mean value found in comparisons with random sequences is shown as a density plot on the grey scale from white (highest SD values) to black (lowest SD values). Although several tones of the grey scale are available, for clarity only four representative values are shown. Reprinted from reference 12a with permission from Cambridge University Press.

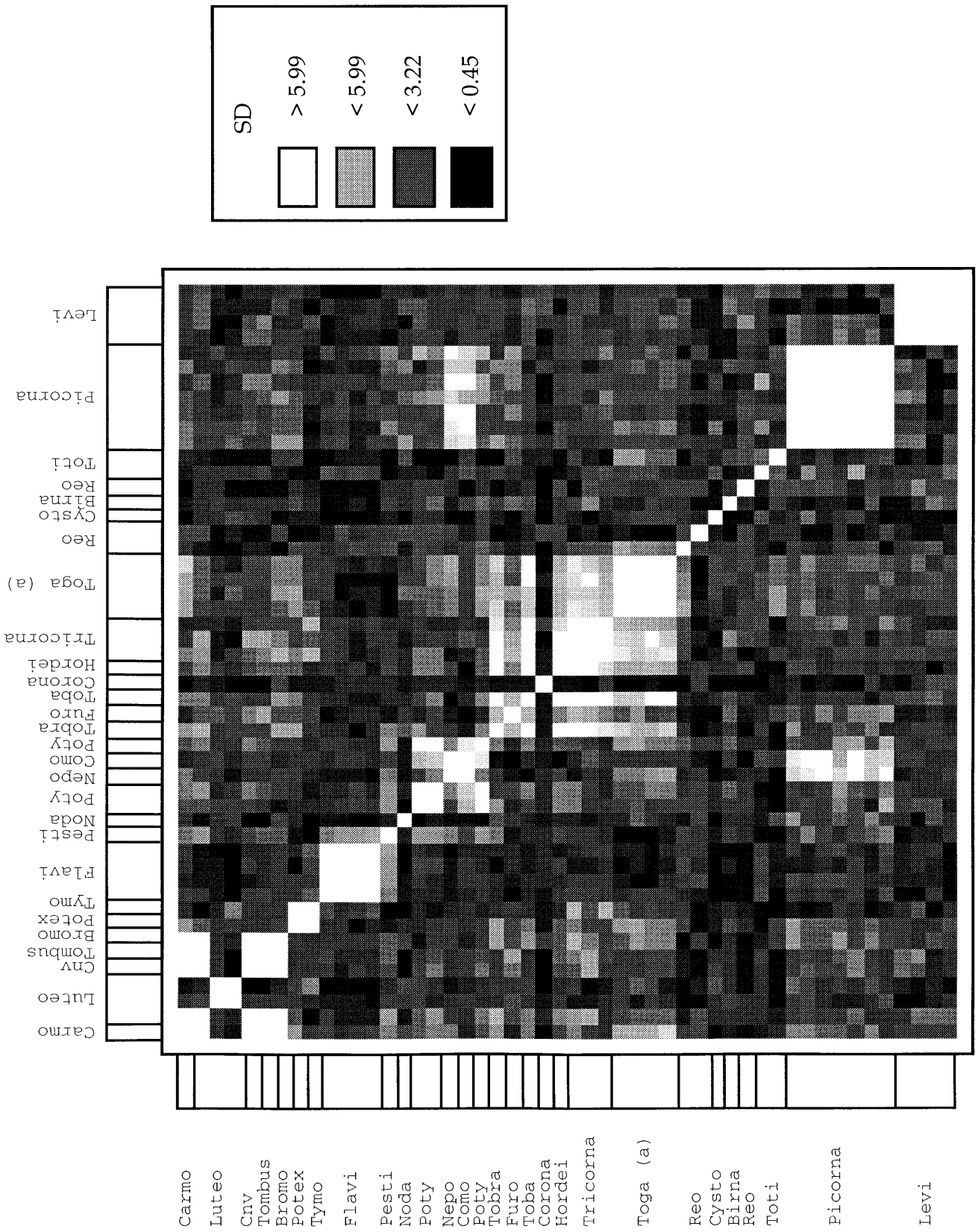

FIG. 2. Monte Carlo simulation density plots for 50 RdRp from Bruenn (5).

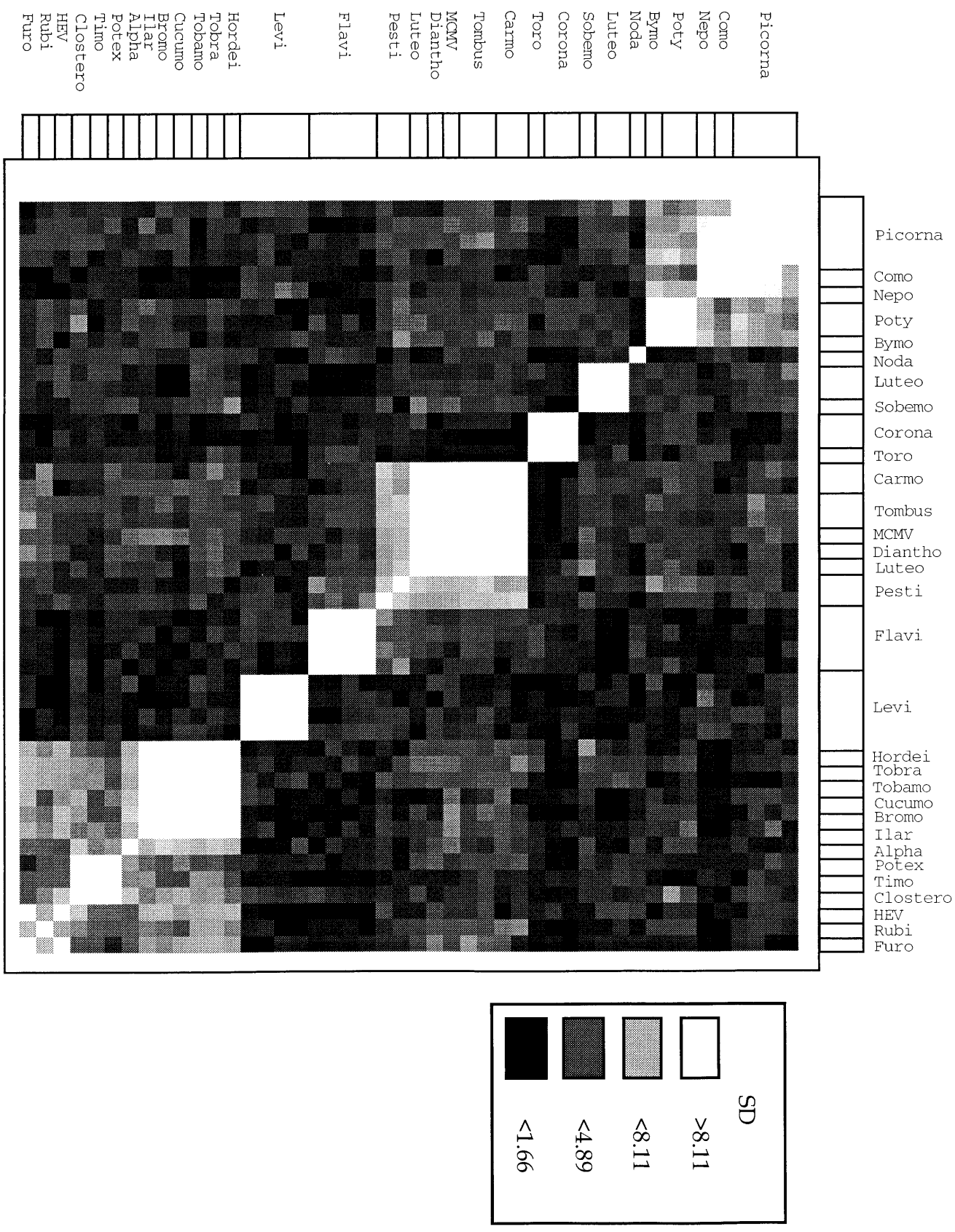

FIG. 3. Monte Carlo simulation density plots for 46 RdRp from Koonin (13).

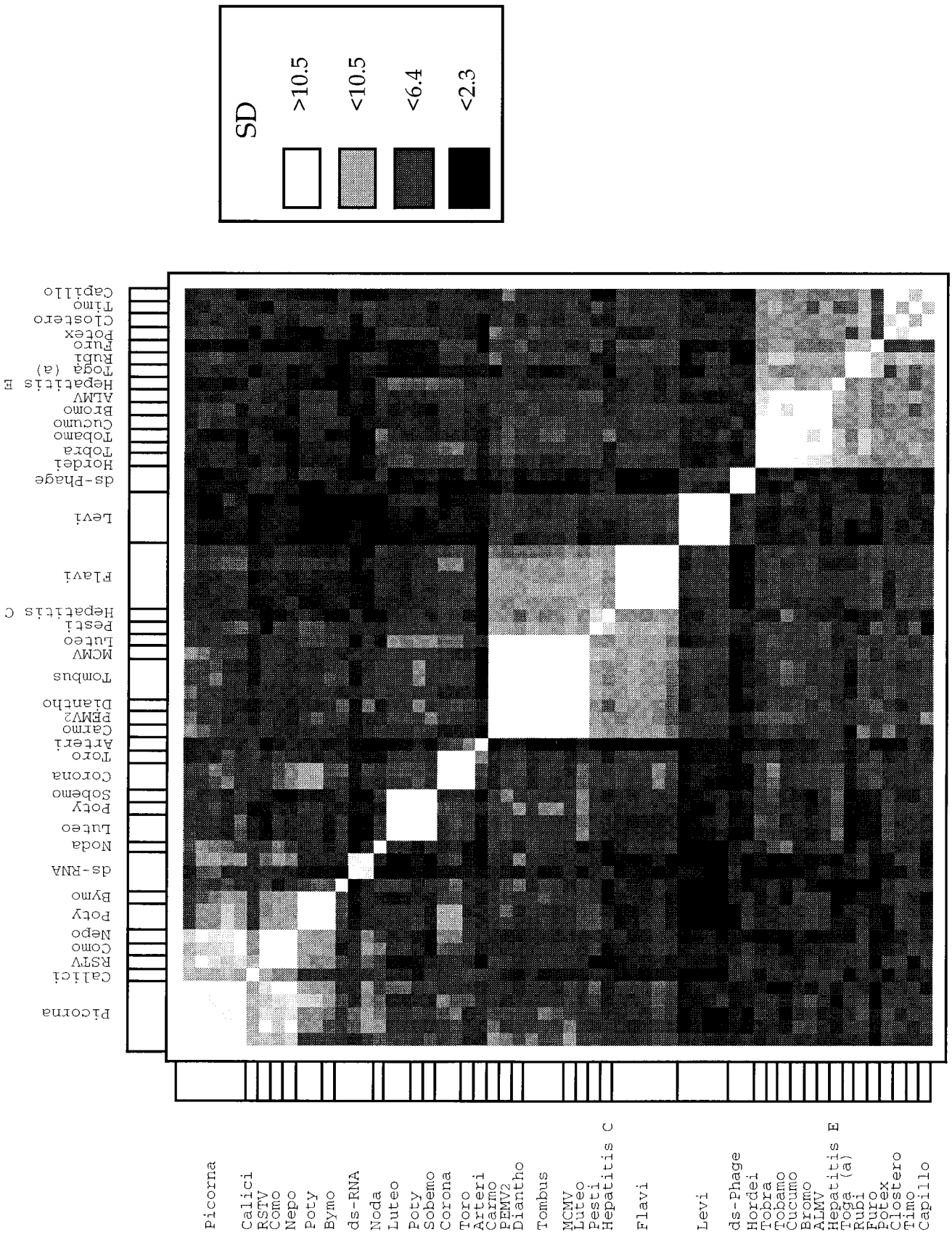

FIG. 4. Monte Carlo simulation density plots for 59 RdRp from Koonin and Dolja (14). Reprinted from reference 12a with permission from Cambridge University Press.

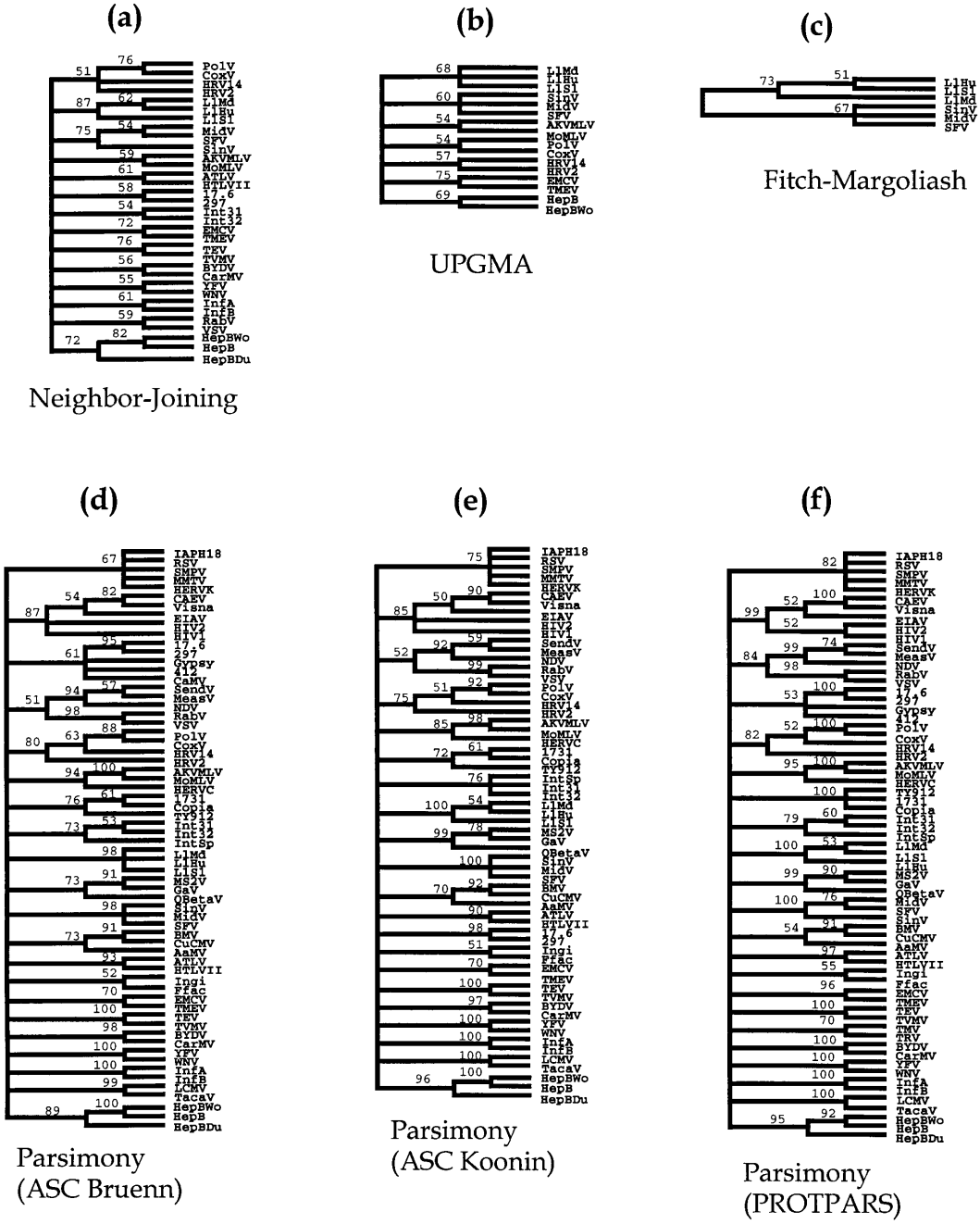

FIG. 5. Unrooted majority rule consensus bootstrap trees for 40 RdRp and 40 RT from Poch et al. (18). Results from the following methods are presented: neighbor joining (a), UPGMA (b), Fitch-Margoliash (c), parsimony with amino acid similarity classes (ASC) defined by Bruenn (5) (d), parsimony with amino acid similarity classes defined by Koonin (13) (e), and parsimony with the PROTPARS program (f). Bootstrap analysis and consensus calculations were done as described in the text. Only taxa which show associations with greater than 50% bootstrap support are shown. Abbreviations of viruses and other agents used in the phylogenetic analysis are given in Table 1. In all cases the leftmost horizontal branch (the tree backbone) has no bootstrap support and is shown only for graphic purposes. Horizontal branch lengths are not drawn to scale.

tyviruses (around 60%) were among the groups supported in the parsimony bootstrap analysis. Overall, however, there was no resolution for this data set beyond the previously recognized groups and the trees obtained using the distance matrix methods are even less resolved (and no branches at all in the Fitch-Margoliash tree could be resolved). In particular, it is striking that hepatitis C virus RdRp clustered with the flavivirus RdRp more than 50% of the time in the parsimony analysis yet in none of the distance-based trees.

For the Koonin data set, shown in Fig. 7, a similar lack of resolution with distance matrix trees was observed, especially for the UPGMA (Fig. 7b) and Fitch-Margoliash (Fig. 7c) methods. However, the neighbor-joining method (Fig. 7a) resolved a subtree consisting of the carmo-, tombus-, maize chlorotic mottle, diantho-, and luteovirus groups, including barley yellow dwarf virus. The same subtree was also found with varying degrees of support in the three parsimony bootstrap trees (Fig. 7d and e). The level of resolution of the parsimony

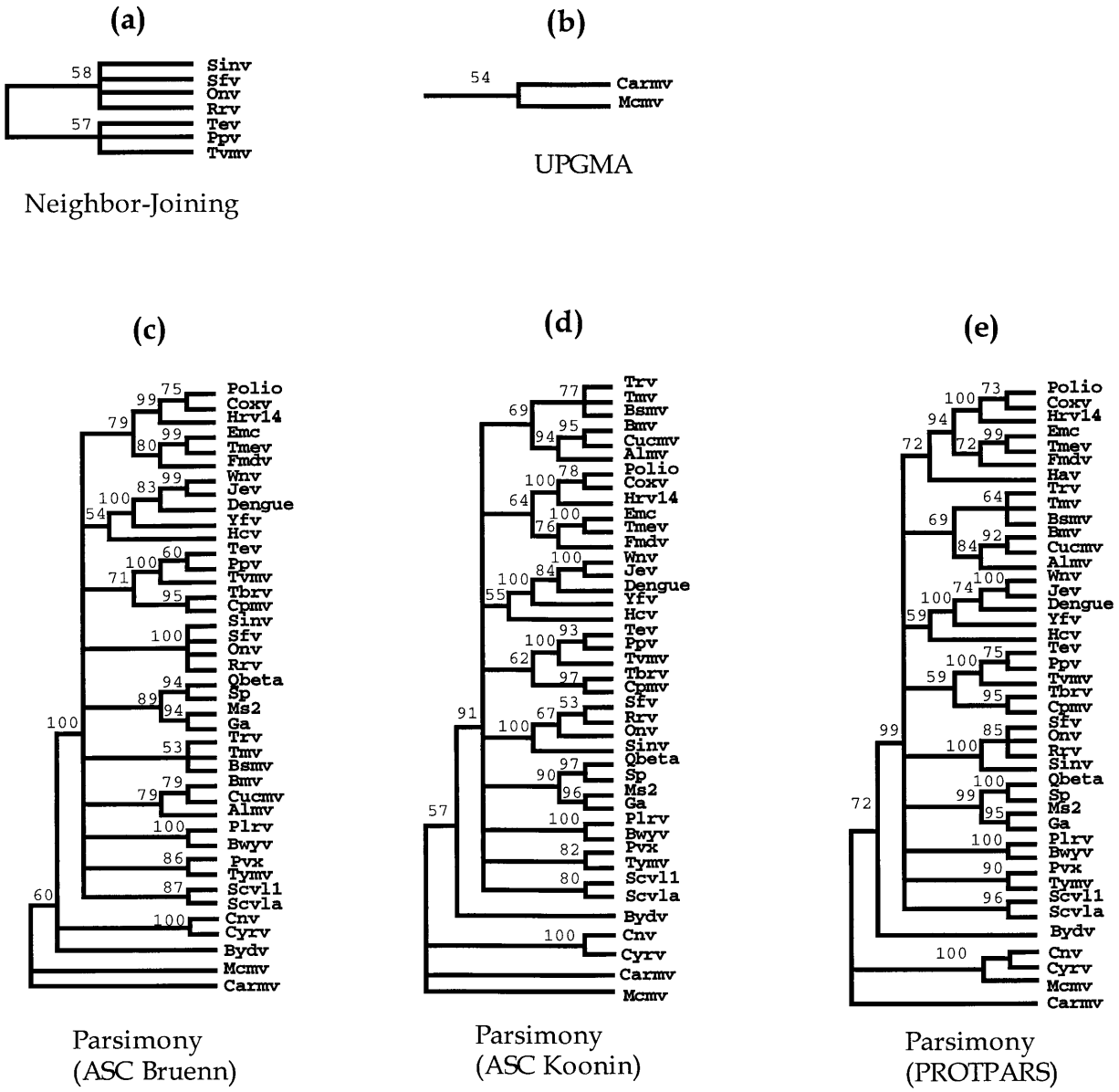

FIG. 6. Majority rule consensus bootstrap trees for 50 RdRp from Bruenn (5). The Fitch-Margoliash tree was not included, since no node could be resolved with more than 50% bootstrap support.

trees for this data set is considerably higher than that of the previous two data sets, and a putative supergroup III (13) was supported 80% of the time (Fig. 7f) although there was no consistent branching order within this group. No support for the other higher-level groupings was obtained. By reducing this data set to regions around the critical residues in the RdRp core (i.e., the Koonin and Dolja data set), no considerable improvement was obtained, as indicated in Fig. 8. In this case, the putative supergroup III lost resolution in the UPGMA bootstrap tree (Fig. 8a), and its support by parsimony did not improve considerably.

In summary, for all four data sets, the 50% majority rule parsimony trees were more resolved than those reconstructed by distance methods, perhaps because of the accumulation of several equally parsimonious solutions with minimal topological differences for a single bootstrap data set. For all tree-building methods on each data set, the bootstrap was unable to give support to the existence of clusters of viral families forming distinct supergroups, with the possible exception of supergroup III.

Phylogenetic signal analysis. On the basis of the results of the bootstrap analysis, a number of better supported viral subtrees were obtained. Representative taxa from these subtrees were extracted, and the strengths of the phylogenetic signals between them were determined. The following subsets of up to nine taxa were used in this analysis (the abbreviations are spelled out in Table 1).

(i) For the Poch et al. data set, the following four sets of taxa were chosen: (i) HepB, HIV2, Copia, Int31, QBetaV, TVMV, SFV, YFV, and BTV; (ii) HepB, HIV2, Copia, Int31, QBetaV, TVMV, SFV, YFV, and VSV; (iii) HepB, HIV2, Copia, QBetaV, TVMV, SFV, YFV, BTV, and FSV; and (iv) HepB,

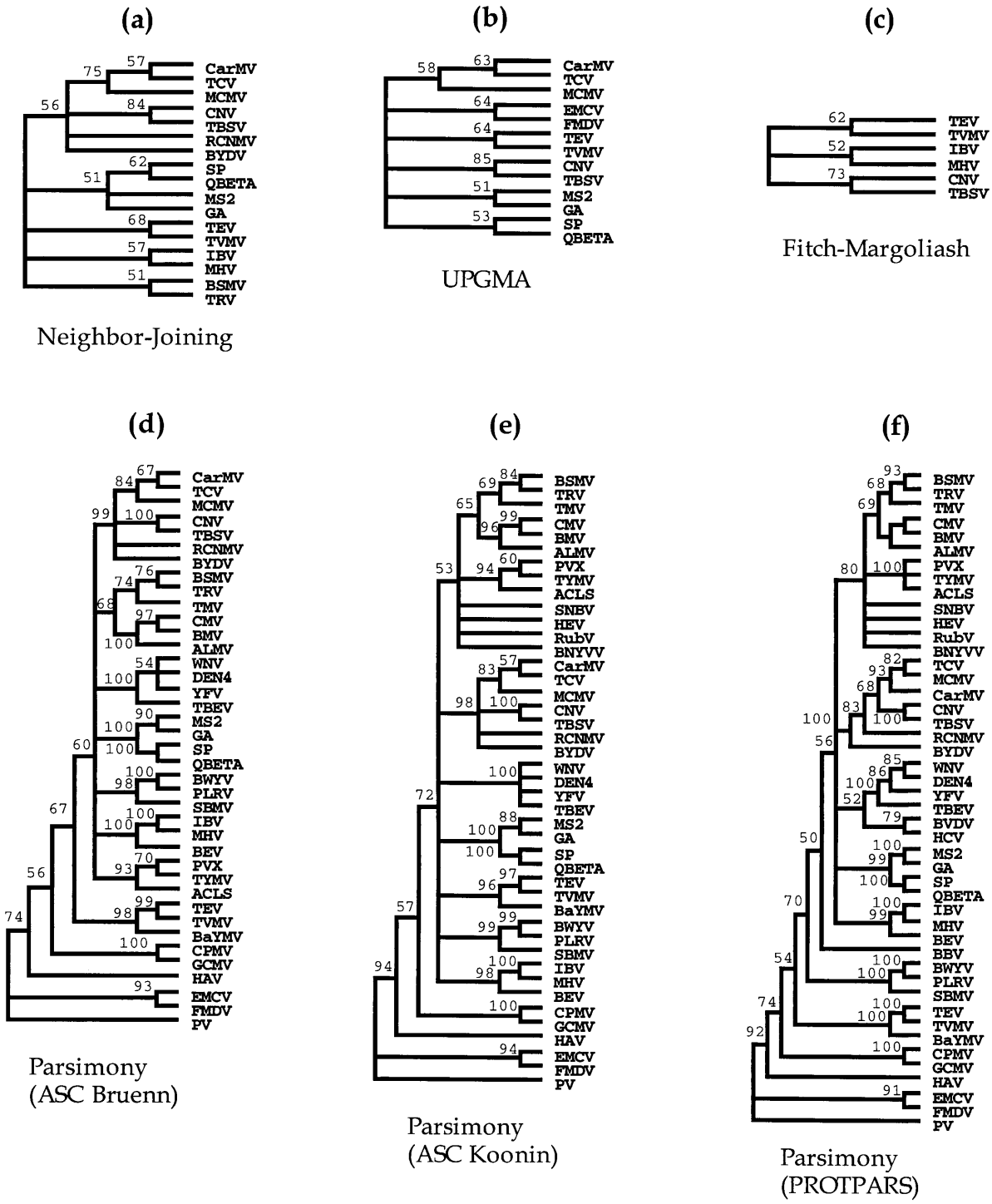

FIG. 7. Majority rule consensus bootstrap trees for 46 RdRp from Koonin (13). Only taxa which show associations with greater than 50% bootstrap support are shown.

CaMV, Copia, Int31, QBetaV, TVMV, SFV, YFV, and VSV. The g1 values obtained were (i) -0.198 ± 0.017, (ii) -0.160 ± 0.017, (iii) -0.230 ± 0.017, and (iv) -0.054 ± 0.017, all of which are indicative of left skew. However, the g1 distribution from 30 equivalent random data sets showed that values less than -0.260 occurred with a frequency of 6.7% along with a minimum g1 value of -0.310. This means that the level of skewness observed in the Poch et al. data set could be obtained by chance within a data set containing no phylogenetic structure.

(ii) For the Bruenn alignment the following data sets were analyzed: (i) Carmv, Tymv, Bbv, Bnyvv, Rot, Phi6, Btv, Polio, and Qbeta; (ii) Wn, Cpmv, Mhv, Cucmv, Onv, Reo, Ibdv, Btv, and Scvla; (iii) Wn, Plrv, Mhv, Bbv, Onv, Rot, Ibdv, Polio, and

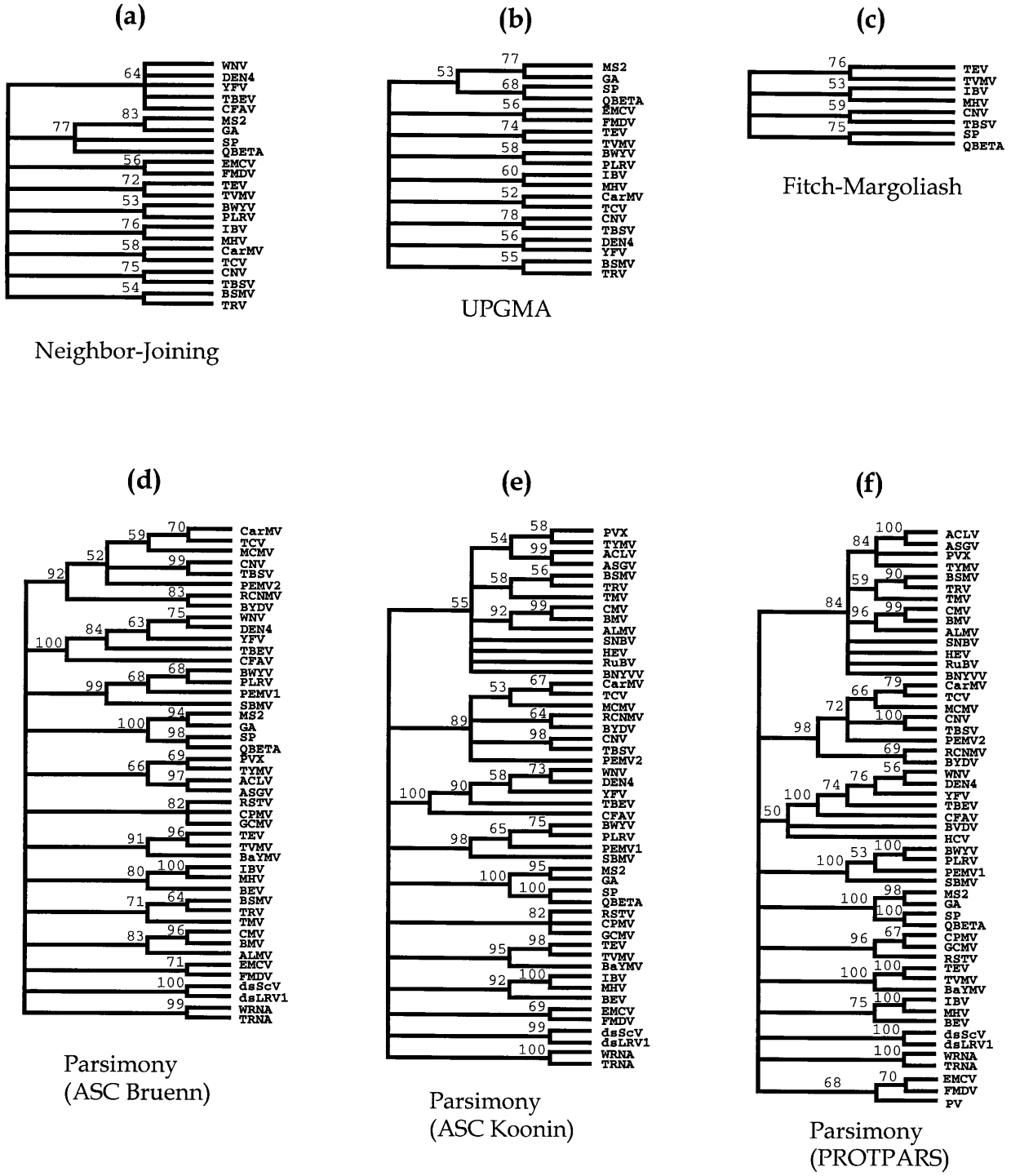

FIG. 8. Majority rule consensus bootstrap trees for 59 RdRp from Koonin and Dolja (14). Only taxa which show associations with greater than 50% bootstrap support are shown.

Sclva; and (iv) Qbeta, Btv, Phi6, Reo, Onv, Bbv, Cucmv, Tymv, and Cpmv. The g1 values obtained were (i) 0.175 ± 0.017, (ii) 0.002 ± 0.017, (iii) 0.192, and (iv) −0.199 ± 0.017. The g1 distribution from the random data sets showed values of <-0.200, with a frequency of 13.3% and a minimum value of −0.590.

(iii) Three data subsets were taken from the Koonin alignment: (i) PV, CPMV, BBV, MHV, CarMV, WN, QBETA, PLRV, and BNYVV; (ii) CPMV, BBV, MHV, CarMV, WN, QBETA, PLRV, and BNYVV; and (iii) TYMV, CPMV, BBV, MHV, CarMV, WN, QBETA, PLRV, and BNYVV. These resulted in g1 values of (i) −0.084 ± 0.017, (ii) 0.041 ± 0.062, and (iii) −0.094 ± 0.017. The g1 distribution from 30 random data sets showed values of <-0.114, with a frequency of 7% and a minimum g1 of −0.360.

(iv) For the modified Koonin and Dolja data set, two subsets were chosen: (i) PV, CPMV, BBV, MHV, PLRV, CarMV, WN, QBETA, and BNYVV; and (ii) TYMV, CPMV, BBV, MHV, PLRV, CarMV, WN, QBETA, and BNYVV. The values for g1 were (i) 0.053 ± 0.017 and (ii) -0.246 ± 0.017. The g1 distribution from the random data sets showed values of <-0.252, with a frequency of 3% and a minimum value of -0.830.

In summary, the g1 values for representative RdRp sequences from each data set did not depart significantly from the random expectation, indicating that there is a lack of reliable phylogenetic structure between distantly related RdRp sequences.

Random tree-length distributions. As described above, representative taxa from distinct subtrees were chosen. However, since for this type of analysis it is not necessary to determine the tree lengths for all possible trees, a larger number of taxa were included. From the Poch et al. alignment 10 representative RdRp were chosen: HepB, HIV2, Copia, Int31, QBeta V, TVMV, SFV, YFV, BTV, and VSV. The most parsimonious tree (552 steps) was longer than the shortest tree from the randomly generated data (550 steps) and not significantly shorter than the average random tree length (99% confidence interval [CI] = 559.857 ± 10.25). For the Bruenn data set, 18 RdRp were chosen: Carmv, Plrv, Tymv, Wnv, Bbv, Cpmv, Bnyvv, Mhv, Cucmv, Onv, Rot, Reo, Phi6, Ibdv, Btv, Scvla, Polio, and Qbeta. In this case the most parsimonious tree (3,954 steps) was again longer than the shortest random tree (3,930 steps) and not significantly shorter than the average random tree length (99% CI = $3{,}962.264 \pm 19.99$). Ten representative RdRp were chosen from the Koonin alignment: PV, CPMV, BBV, MHV, CarMV, WNV, QBETA, PLRV, TYMV, and BNYVV. In this case the most parsimonious tree (2,089 steps) had the same length as the shortest random tree but was not shorter than the average random tree (99% CI = $2{,}103.903 \pm 14.96$). Finally, a subset of 10 taxa were chosen from the modified Koonin and Dolja alignment: BSMV, WNV, QBETA, CPMV, BBV, PLRV, dsScv, RuBV, WRNA, and BNYVV. The most parsimonious tree in this case (898 steps) was only three steps shorter than the shortest tree from the random data (901 steps) and not significantly shorter than the average random tree length, indicative of an unreliable data set for phylogenetic reconstruction (99% CI = 912.578 ± 11.46).

In summary, the most parsimonious trees from representative subsets of taxa are not significantly shorter than those reconstructed from randomly generated data sets. This corroborates the analysis of g1 statistics in showing that there is an absence of reliable phylogenetic signal in the RdRp sequence data.

DISCUSSION

RdRp as a phylogenetic marker. The results presented in this work clearly show that RdRp sequences cannot be used to construct a single phylogenetic tree including all RNA viruses. This is because of a lack of both basic sequence similarity and reliable phylogenetic signal and provides an explanation for why different sequence alignments and tree reconstruction methods produce incongruent phylogenies and consequently why published trees do not agree. Even when data sets with different sizes were used, with some alignments consisting of little more than conserved motifs (for example the Poch et al. data set) there was no improvement in phylogenetic resolution even though there was an increase in the Monte Carlo SD values (although off-diagonal values remained low). This was clearly indicated by the lack of bootstrap support and phylogenetic signal, even between the most frequently found viral groups.

In terms of virus classification, our most important finding was the paucity of support for the clustering of virus families into 11 groups (13) or three larger supergroups, with the possible exception of a putative supergroup III (10, 13, 14, 22). Supergroup III was only obtained by the parsimony method, and even this may be an artifact, as parsimony does not correct for multiple superimposed substitutions and therefore, at this level of divergence, does not reliably disentangle changes due to common ancestry from convergent or parallel changes.

The most convincing evidence in favor of the supergroups is that members appear to carry homologous genes arranged in the same or a variable order and that similar mechanisms are used to express some of these genes (10, 13, 14). However, several of these genes were identified from alignments of apparently conserved motifs, as in the case of the polymerases. For example, an alignment of 43 viral helicases indicated that 95 residues in seven conserved motifs, encompassing the nucleoside triphosphate binding pattern (11, 21), appear to be shared among members of putative supergroups I and II (13). However, for supergroup III a different set of 45 residues in three motifs was identified (13). Therefore, no all-inclusive alignment can be obtained from helicase sequences, undermining their validity as phylogenetic markers across all RNA viruses. Furthermore, the recognition of these motifs relies on position homology and accurate alignment, and it has been shown that for sequences with SD values approaching 0, alignment accuracy is predictably poor (1). Consequently, even assuming that some colinear motifs are unambiguously shared by several polymerase sequences (18), they were not sufficient to improve SD scores on data sets which maximize their contribution (e.g., Poch et al.), nor did they improve the phylogenetic signal of the data at higher nodes. It is also noticeable that even within virus families, such as the family *Flaviviridae*, which currently includes hepatitis C virus, pestiviruses, and flaviviruses, no unequivocal evidence for common ancestry was obtained. Incongruent phylogenetic trees were also obtained with the RdRp and helicase sequences from two GB agent hepatitis viruses and the other members of the family *Flaviviridae* (17). We have reexamined these data sets (a 300-residue-long helicase alignment and a 298-residue-long RdRp alignment) using the Monte Carlo method. On average, SD values greater than 6 were observed in comparisons involving the helicases but much lower levels were found with the RdRp sequences (data not shown). Therefore, even when focusing on the virus family level it is often not possible to resolve phylogenetic relationships. In conclusion, we feel that it is more appropriate to present the evolutionary relationships between RNA viruses as a set of distinct subtrees, the links between which are unclear, rather than as a single and resolved phylogenetic tree.

The origin of RNA polymerases. The lack of phylogenetic signal at higher taxonomic levels also raises the question of whether there was a common ancestor for all RdRp and RT and even for the RdRp alone. In view of the lack of support for monophyly of RNA polymerases from the data, it could be argued that the RNA polymerase function arose independently during the radiation of distinct lineages of viruses and transposable elements. The lack of conservation in primary sequence and size among polymerases suggests that the requirements for polymerase function can be fulfilled by diverse means, supporting the notion of multiple origins. Convergent evolution could then explain the presence of many of the conserved motifs among paraphyletic RdRp, an example of which is the conservation of RdRp-like motifs in components of telomerases (15). In this case, a taxonomy based on a single

phylogenetic tree would have little evolutionary meaning. On the other hand, it could be that the RdRp are monophyletic and that several processes, including extreme sequence divergence caused by the fast evolutionary rates which characterize RNA viruses and the extinction of ancestral lineages, could have erased the phylogenetic signal which should link virus families.

In summary, it was found that our analysis of several published data sets did not allow the inclusion of all RNA viruses in a single phylogenetic framework as proposed by previous authors. This can be explained by a lack of sequence similarity and loss of phylogenetic information, causing phylogenetic methods to generate unresolved and minimally informative trees with no explanatory power at the higher taxonomic levels. Therefore, we suggest that new taxonomic categories based on molecular systematic analysis should be considered only for highly supported groups. It remains to be seen if additional sequence information and improved alignments of the RdRp and other conserved genes such as the helicases will eventually produce more robust phylogenetic trees from divergent viral genomes.

ACKNOWLEDGMENTS

We thank David C. Krakauer for helping with the data analysis and presentation, Andrew Rambaut and Liz Cowe for support with the computer analysis, and Ian Cooper for valuable discussion.

This work was supported by grants from The Royal Society and The Wellcome Trust.

REFERENCES

1. **Barton, G. J.** 1990. Protein multiple sequence alignment and flexible pattern-matching. Methods Enzymol. **183:**403–428.
2. **Barton, G. J.** Protein sequence alignment and database scanning. *In* M. J. E. Sternberg (ed.), Protein structure prediction: a practical approach, in press. IRL Press, Oxford (available at http://geoff.biop.ox.ac.uk/).
3. **Barton, G. J., and M. J. E. Sternberg.** 1987. A strategy for the rapid multiple alignment of protein sequences—confidence levels from tertiary structure comparisons. J. Mol. Biol. **198:**327–337.
4. **Bishop, D. H. L.** 1994. Virus taxonomy: alive and well! Soc. Gen. Microbiol. Q. **21:**36–38.
5. **Bruenn, J.** 1991. Relationships among the positive-strand and double-strand RNA viruses as viewed through their RNA-dependent RNA polymerases. Nucleic Acids Res. **19:**217–225.
6. **Dolja, V. V., and J. C. Carrington.** 1992. Evolution of positive-strand RNA viruses. Semin. Virol. **3:**315–326.
7. **Eickbush, T. H.** 1994. Origin and evolutionary relationships of retroelements, p. 121–157. *In* S. S. Morse (ed.), The evolutionary biology of viruses. Raven Press, New York.
8. **Felsenstein, J.** 1993. PHYLIP (phylogeny inference package), version 3.5c. Department of Genetics, University of Washington, Seattle.
9. **Feng, D. F., and R. F. Doolittle.** 1985. Aligning amino acid sequences: comparison of commonly used methods. J. Mol. Evol. **21:**112–125.
10. **Goldbach, R., and P. de Haan.** 1994. RNA viral supergroups and evolution of RNA viruses, p. 105–119. *In* S. S. Morse (ed.), The evolutionary biology of viruses. Raven Press, New York.
11. **Gorbalenya, A. E., and E. V. Koonin.** 1993. Helicases. Amino acid sequence comparisons and beyond. Curr. Opin. Struct. Biol. **3:**419–429.
12. **Hillis, D. M.** 1991. Discriminating between phylogenetic signal and random noise in DNA sequences, p. 278–294. *In* M. M. Miyamoto and J. Cracraft (ed.), Phylogenetic analysis of DNA sequences. Oxford University Press, Oxford.

12a. **Holmes, E. C., E. A. Gould, and P. M. de A. Zanotto.** An RNA virus tree of life? *In* D. M. Roberts, P. Sharpe, G. Alderson, and M. Collins (ed.), Evolution of microbial life. Society for General Microbiology Symposium 54, in press. Cambridge University Press, Cambridge.

13. **Koonin, E. V.** 1991. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. J. Gen. Virol. **72:**2197–2206.
14. **Koonin, E. V., and V. V. Dolja.** 1993. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit. Rev. Biochem. Mol. Biol. **28:**375–430.
15. **Lundbald, V., and E. H. Blackburn.** 1990. RNA-dependent polymerase motifs in EST1: tentative identification of a protein component of an essential yeast telomerase. Cell **60:**529–530.
16. **Manly, B. F. J.** 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, London.
17. **Muerhoff, A. S., T. P. Leary, J. N. Simons, T. J. Pilot-Matias, G. J. Dawson, J. C. Erker, M. L. Chalmers, G. G. Schlauder, S. M. Desai, and I. K. Mushahwar.** 1995. Genomic organization of GB viruses A and B: two new members of the *Flaviviridae* associated with GB agent hepatitis. J. Virol. **69:**5621–5630.
18. **Poch, O., I. Sauvaget, M. Delarue, and N. Tordo.** 1989. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. EMBO J. **8:**3867–3874.
19. **Stuart, A., and K. Ord.** 1994. Kendall's advanced theory of statistics, 6th ed., vol. I. Distribution theory. Edward Arnold, London, United Kingdom.
20. **Swofford, D. L.** 1993. Phylogenetic analysis using parsimony (PAUP), version 3.1.1. University of Illinois, Champaign.
21. **Walker, J. E., M. Sarasate, M. J. Runswick, and N. J. Gay.** 1982. Distantly related sequences in a- and b-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J. **1:**945–951.
22. **Ward, C. W.** 1993. Progress towards a higher taxonomy of viruses. Res. Virol. **144:**419–453.
23. **Xiong, Y., and T. H. Eickbush.** 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9:**3353–3362.